Go Main Menu | Power Search | List all topics | List all methods

Story Name:
US Crime
Story Topics:
Social science
Datafile Name:
US Crime
Methods:
Collinearity , Correlation , Causation , Lurking variable , Regression
Abstract:
These data are crime-related and demographic statistics for 47 US states in 1960. The data were collected from the FBI's *Uniform Crime Report* and other government agencies to determine how the dependent variable crime rate (R) depends on the other variables measured in the study.

We encounter many problems analyzing these data by regression because some predictor variables are highly correlated. For example, Ex0 and Ex1, which measure police expenditures in consecutive years, have a correlation of .99. Wealth (W) and income inequality (X) are also highly correlated, as are U1 and U2, which measure unemployment in two different age groups. When predictor variables are highly correlated, the model is said to be nearly collinear. The result is that our estimated coefficients are unstable; removing one variable from the model may cause the results for the other variables to change dramatically.

In addition, the causal relationship between Ex0 (expenditures in 1960) and crime rate is unclear. Do increased expenditures affect the crime rate, or does the crime rate motivate an increase in expenditures?

In one possible analysis, predictors are removed from the model until only the 5% significant predictors Age, Ed, U2, X, and Ex0 remain. The results of this model are in Figure 1. This model demonstrates that it is important to look at the direction of the coefficients. From these coefficients, it appears that more education and police expenditures increase the crime rate. Perhaps there is another variable, a "lurking variable" not collected with these data, which causes both education and crime rate to increase together.

This data set is a good example of what can go wrong in a regression analysis.
Image:
Results for a possible model for these data

Dependent variable is: **R**
No Selector
48 total cases of which 1 is missing
R squared = 73.0% R squared (adjusted) = 69.7%
s = 21.30 with 47 - 6 = 41 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	50205.6	5	10041.1	22.1
Residual	18603.6	41	453.747	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-524.374	95.12	-5.51	≤ 0.0001
Age	1.01982	0.3532	2.89	0.0062
Ed	2.03077	0.4742	4.28	0.0001
U2	0.913608	0.4341	2.10	0.0415
X	0.634926	0.1468	4.32	≤ 0.0001
Ex0	1.23312	0.1416	8.71	≤ 0.0001